1 Описание

Мечи залива работорговцев

После оглушительного успеха в освобождении Астапора, Миэрина и Юнкая от власти работорговцев Дейенерис Бурерожденная открыла себе доступ к Летнему морю, а следовательно – путь в Вестерос. Для ведения войны с Семью Королевствами нужно оружие, я для оружия нужна сталь. Нет никаких сомнений в кузнечном искусстве Безупречных, однако поставщики стали не столь надежны. Два основных поставщика стали — это Westeros Inc. и Harpy & Co. На протяжении нескольких месяцев мы закупаем сталь у обеих компаний, и каждая из них предлагает ощутимую скидку при заключении эксклюзивного договора на поставку. Советник королевы Тирион Ланнистер знает о твоем умении принимать взвешенные рациональные решения и просит помощи в объективном решении вопроса о том, с какой из компаний следует заключить эксклюзивный договор на поставку стали. У Тириона есть записи о производстве мечей каждым из кузнецов-безупречных, а также данные о количестве сломанных мечей в каждый из месяцев ведения боевых действий.

2 Исходные данные

CSV-файл с данными о производстве оружия и количестве единиц сломанного оружия за каждый месяц каждым из кузнецов.

3 Постановка задачи

Необходимо провести разведывательный анализ данных с целью ответа на вопрос: "С каким из поставщиков стали следует заключить договор?"

Основные моменты:

1.Код должен быть оформлен в виде Python notebook в файле analysis.ipynb.

2.Результаты анализа и ваши выводы должны быть оформлены в виде презентации средствами latex и beamer.

3.Исходный код презентации должен быть в файле presentation.tex, ее отрендеренный вариант — в файле presentation.pdf.

4.Описание подхода (т.е. стандартное требование ко всем заданиям) в формате pdf.

4 Теоретическая справка

В статистике разведывательный анализ данных (EDA) — это подход к анализу наборов данных для обобщения их основных характеристик, зачастую с помощью визуальных методов. Статистическая модель может использоваться или нет, но в первую очередь EDA предназначена для того, чтобы

увидеть, что данные могут показать нам, помимо формальных задач моделирования или проверки гипотез. Джон Тьюки продвигал разведывательный анализ данных, чтобы побудить статистиков исследовать данные и, возможно, формулировать гипотезы, которые могли бы привести к сбору новых данных и новым экспериментам. EDA отличается от анализа начальных данных (IDA), который более узко фокусируется на проверке допущений, необходимых для подбора модели и проверки гипотез, а также на обработке пропущенных значений и преобразованиях переменных по мере необходимости. EDA включает в себя IDA.

Задачами EDA являются:

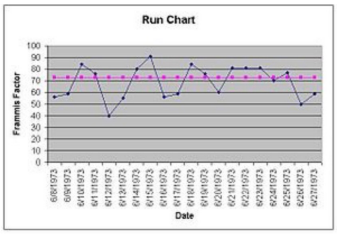
1. Предложение гипотез о причинах наблюдаемых явлений
2. Оценка предположений, на которых будет основываться статистический вывод
3. Подтверждение выбора соответствующими статистическими инструментами и методами
4. Обеспечение основы для дальнейшего сбора данных с помощью опросов или экспериментов

Многие методы EDA были применены в интеллектуальном анализе данных, а также в аналитике больших данных.

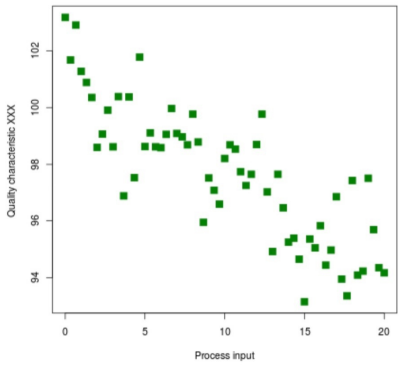
Существует ряд инструментов, которые полезны для EDA, но EDA характеризуется скорее подходом, чем конкретными методами.

Некоторые из типичных графических методов, используемых в EDA:

1. **Диаграмма размаха** *(box-and-whiskers diagram or plot, box plot)* — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей. (Мы не будем брать этот метод потому что медиана не информативна для нас, а он основывается на ней)
2. **Гистограмма**
3. **График прогона**, также известный как график последовательности прогонов, представляет собой график, который отображает данные наблюдений во временной последовательности. Часто данные показывают производительность некоторого процесса, поэтому это форма линейного графика.



4. **Диаграмма рассеяния** (также **точечная диаграмма**, *scatter plot*) — математическая диаграмма, изображающая значения двух переменных в виде точек на декартовой плоскости. (Не очень удобно)



5. **Параллельные координаты** являются распространенным способом визуализации многомерной геометрии и анализа многомерных данных. (У нас нет многомерных данных -> тоже не подходит)

6. **Отношение шансов** — характеристика, применяемая для количественного описания тесноты связи признака А с признаком Б в некоторой статистической популяции. (Тоже довольно бесполезно для нашей задачи)

5 Подход к решению

При решении задачи будем использовать графики прогона, так как этот метод отображает данные во временной последовательности, а нам как раз нужно анализировать информацию в некоторые промежутки времени.

Используемые библиотеки:

- csv
- matplotlib

Первая необходима для считывания данных из файла .csv. Это делается с помощью метода DictReader, который считывает данные до заданного разделителя.

С помощью методов библиотеки matplotlib отрисовываем графики, анализируя которые приходим к выводу о том, какая компания подходит больше. Графики и их анализ подробно описаны в файле presentation.pdf.

6 Работу выполнили

Евенко Алеся, Ефарова Дарья, Коротчук Анастасия - совместная работа над кодом, презентацией и README